Gene Homology Explorer

Proposal

Josh Goodman & Mark Green

Spring 2023

Contents

${toc}

1 Abstract

2 Introduction

In the field of human genetic research, model organisms play a crucial role in helping to decipher functional mechanisms, disease mechanisms, variant impact, and many other aspects of genes[1]. Researchers in this field of study rely on previously published data in their organism of interest and also related organisms to discover as much information as possible. A geneticist studying the KRAS gene in humans might look for studies on related genes in mice or rats before designing experiments or looking for drug targets. These related genes are called *orthologs*. Orthologs are homologous genes that are the result of a speciation event[2]. In other words, a gene in one species that is directly, but possibly distantly, related to a gene in another species over an evolutionary time period. *Paralogs*, genes that are the result of a duplication event within a species, can also be used for this same purpose (Figure 1).

[1] needs citation

[2] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38. doi: 10.1146/annurev.genet.39.073003.114725. PMID: 16285863.

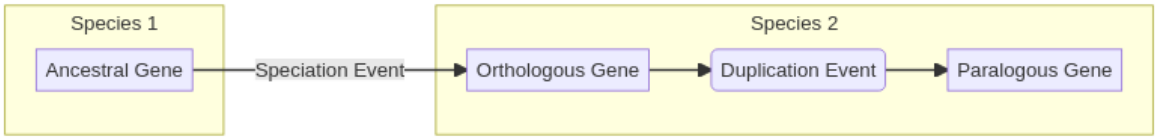


Figure 1: Origin of orthologous and paralogous genes.

2.1 Motivation

The exact definition of what constitutes an ortholgous or paralogous pair of genes has been an active area of study for decades[3]. Over this time, many orthology prediction algorithms have been developed, making it difficult for researchers to select one over the other. To address this issue a meta-orthology tool called DIOPT[4] was developed by the Perrimon lab at Harvard Medical School. DIOPT takes the approach of aggregating as many orthology and paralogy algorithm prediction results as possible and presenting all to the end user when a search for one or more genes is conducted. Each homologous pair of genes is scored according to the number of algorithms that have predicted their evolutionary relationship. The tool allows users to enter one or more genes and view results in a tabular format.

While useful, this functionality fails to convey the relationships between the genes being queried within a species, relationships to orthologous genes in other species, and paralogous genes in a visual manner. Herein, we propose the development of a network visualization tool that will allow researchers to explore these relationships, filter based on species, algorithm scores, or other attributes, and easily link out to primary source databases for additional information.

2.1.1 Related Work

To date, the presentation of results from DIOPT have been limited to tabular HTML results or downloadable tab separated files[5] [6].

3 Project Proposal

Herein, we propose the development of a network visualization tool that will allow researchers to explore these relationships, filter based on species, algorithm scores, or other attributes, and easily link out to primary source databases for additional information. Below are discussed details of the technical approach to accomplish this task and a project Road Map to outline key milestones and goals throughout the project duration.

3.1 Technical Approach

TODO: Temporary notes on ideas for this section

- Data acquisition, processing, and analysis
  - Python
- Creation of a data warehouse for storing graph information
  - NoSQL: DuckDB, SQLite, MongoDB, JSON
  - SQL: PostgreSQL

[3] needs citation

[4] Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357. doi: 10.1186/1471-2105-12-357. PMID: 21880147; PMCID: PMC3179972.

[5] Alliance of Genome Resources Consortium. Harmonizing model organism data in the Alliance of Genome Resources. Genetics. 2022 Apr 4;220(4):iyac022. doi: 10.1093/genetics/iyac022. PMID: 35380658; PMCID: PMC8982023.

[6] Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, Larkin A, Matthews BB, Millburn G, Strelets VB; the FlyBase Consortium. FlyBase: a guided tour of highlighted features. Genetics. 2022 Apr 4;220(4):iyac035. doi: 10.1093/genetics/iyac035. PMID: 35266522; PMCID: PMC8982030.

    - GraphDB: Dgraph, JanusGraph, or Neo4J
- Possible network analysis to add value to homology data
- Web application that is either pregenerated or fetches data from a "live" API to provide data exploration, filtering, and basic analysis
    - Web frameworks
        * Astro
        * Remix
    - Network visualization
        * Cytoscape.js
        * Sigma
        * Vega

3.1.1 Data Sources

Briefly discuss how we plan to acquire the data, from what sources, and what (if any) processing might be required.

- Meta orthologs and paralogs [1]
- Alliance of Genome Resources - Various model organism data (functional, disease associations, etc.) [3].

3.1.2 Data Warehousing and Transformation

No need for a lot of detail here, but I think we should give a high level overview of building a data warehouse. The whys and possible hows.

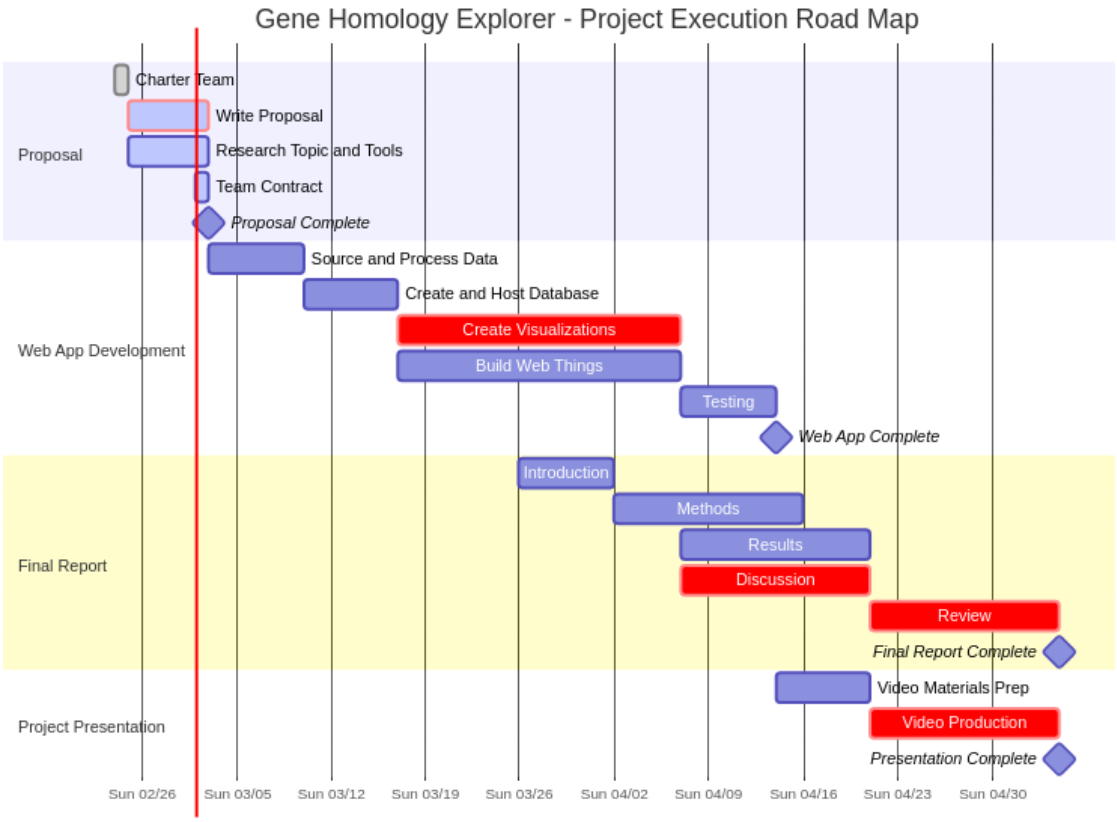
3.1.3 Network Analysis Methods

Any processing and data analysis that we think we might want to do.

3.1.4 Visualization

Talk a little bit about our visualization options. Maybe mention the limitations of each in terms of the network size that they support.

3.2 Project Execution Roadmap



4 Acknowledgments

5 References

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357. doi: 10.1186/1471-2105-12-357. PMID: 21880147; PMCID: PMC3179972.

Wang J, Al-Ouran R, Hu Y, Kim SY, Wan YW, Wangler MF, Yamamoto S, Chao HT, Comjean A, Mohr SE; UDN; Perrimon N, Liu Z, Bellen HJ. MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. Am J Hum Genet. 2017 Jun 1;100(6):843-853. doi: 10.1016/j.ajhg.2017.04.010. Epub 2017 May 11. PMID: 28502612; PMCID: PMC5670038.

Alliance of Genome Resources Consortium. Harmonizing model organism data in the Alliance of Genome Resources. Genetics. 2022 Apr 4;220(4):iyac022. doi: 10.1093/genetics/iyac022. PMID: 35380658; PMCID: PMC8982023.

Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38. doi: 10.1146/annurev.genet.39.073003.114725. PMID: 16285863.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, Larkin A, Matthews BB, Millburn G, Strelets VB; the FlyBase Consortium. FlyBase: a guided tour of highlighted features. Genetics. 2022 Apr 4;220(4):iyac035. doi: 10.1093/genetics/iyac035. PMID: 35266522; PMCID: PMC8982030.